PART 1

INTRODUCTION

Chapter 3. Information Management

David A. Griesmer
Computer Sciences Corporation
Large Lakes Research Station
9311 Groh Road
Grosse Ile, Michigan 48138

To support the modeling efforts of the Lake Michigan Mass Balance Project (LMMBP), large amounts of data were collected and analyzed by a number of State and government agencies and universities (Appendix 1.3.1). Data were collected, analyzed, and sent to the United States Environmental Protection Agency (USEPA) Great Lakes National Program Office (GLNPO) in Chicago, Illinois. GLNPO staff, under the direction of Lou Blume, were responsible for quality assurance (QA) assessment, organization, and consolidation of all data. To facilitate the QA assessment process, a SAS application Research Data Management and Quality Assurance System (RDMQ), developed by Syd Allen, a private contractor, was used to automate the QA process (Sukloff, 1995). RDMQ is a menu-driven SAS program. It has capabilities for loading data, applying quality control checks, adding validity flags, viewing and editing data, producing user-defined tables and graphs, and exporting data in ASCII files. These tasks are performed through a set of menu-driven SAS programs and macros. Data which had been put through the assessment process and approved for release by both GLNPO and the Principal Investigator (PI) were then sent to USEPA, Office of Research and Development (ORD)/National Health and Environmental Effects Research (NHEERL)/Large Lakes and Rivers Forecasting Research Branch (LLRFRB)/Large Lakes Research Station (LLRS) for use by the modeling staff.

1.3.1 Overview of Information Management at the LLRS

Data received from GLNPO were usually in the form of electronic media. Data were typically E-mailed, but sometimes they were downloaded from GLNPO databases or received on CD-ROM. Data were reformatted by GLNPO into a form facilitating entry into database programs at the LLRS. Upon arrival, raw data were copied to the “lmmb” folder on Dave Griesmer’s personal network space (“n:\” drive). In addition, data were imported into one of several Microsoft Access databases in the “\Access_2000\lmmb” folder on Mr. Griesmer’s “n:\” drive. The “n:\” drive was used to facilitate data security because this file space is backed up regularly and is only available to Mr. Griesmer. Data were placed in the Microsoft Access databases to facilitate data review/assessment and later retrieval for the modeling team.

Prior to use, several reviews were done of the data received to look for errors in the data sets. At the LLRS, this review was broken up into two parts. First, an initial review was made to check for completeness of information, to look for transcription errors, programming errors, and formatting errors, and to review comments added by collection and analysis personnel. Second, a review was done by the data users to determine if the data made environmental sense. This type of review was conducted for the open lake, surficial sediment and sediment trap, lower food chain, and fish polychlorinated biphenyl (PCB) data sets. Atmospheric PCB fluxes/loadings and tributary PCB loadings did not go through this review process at the LLRS, but they were assessed by study members assigned with providing loading values. Tributary PCB loading assessment was done by David Hall, U.S. Geological Survey (USGS). All atmospheric PCB loading/concentration data were assessed by Keri Hornbuckle, University of Iowa. The assessment process used by these individuals is unknown.

For data reviewed at LLRS, samples which GLNPO determined had failed the RDMQ QA process were flagged with the value of -9999. GLNPO preserved the values in the data sets that were received and flagged the analytical remark field for that parameter. Flagging these values as -9999 facilitated processing by analytical software such as IDL. Parameter values with analytical remark flags of "INV" (invalid data, as determined by GLNPO QA evaluation), and "NAI" (no result reported - interference) were changed to -9999. Samples with the analytical remark flag of "LAC" (no results reported, laboratory accident) were removed.

Documentation associated with the data was studied. RDMQ data warning fields (RS_NMAND, RS_WARN, RS_UPDAT) were checked to verify that there were no problems flagged by RDMQ which were inadvertently included in the database. Every routine field sample (RFS) and field duplicate (FD#) was checked to verify that a valid station name, sampling date, and depth collection information were included. The value ranges (minimum, maximum, average) for all congeners was checked to look for any obvious errors. Data ranges of all data were also checked for obvious errors. Data were checked to verify units and to confirm whether blank, dilution, or surrogate corrections were done. Quality Control (QC) Coordinator (RECSTAT), station notes (STNNOTES), and record (RECSTAFF) comment fields were checked for comments associated with a sample. All of this information was recorded on a Data Verification Checklist (Appendix 1.3.2). If questions or errors were found, they were referred back to GLNPO for resolution.

Upon completion of this initial data check, readme files were created to describe the data, and the raw data set(s) and readme files were copied to a data archive on the LLRS Unix systems. This archive is located at \usr\lmmbdata on the Alpha workstation named llrssrv2 and is available to modeling staff at the LLRS. Each study has its own directory (LMI0001 - LMI0041) within the lmmbdata archive. PCB data for the LMMBP can be found in directories: LMI0029 (daily gas phase congener, total PCBs, and *trans*-nonachlor for each surface cell in the LMMBP 5 km grid); LMI0032 (particulate and precipitation congener, total PCBs, and *trans*-nonachlor data for eight onshore sampling stations around Lake Michigan and for shipboard sampling from the Lake Guardian); LMI0035 (open lake congener, total PCBs, and *trans*-nonachlor data collected during the eight LMMBP cruises conducted in 1994-1995); LMI0036 (phytoplankton, zooplankton, *Mysis* and *Diporeia* congener, total PCBs, and *trans*-nonachlor data collected during eight LMMBP cruises); LMI0037 (forage and predator (lake trout, coho salmon) fish congener, total PCBs, and *trans*-nonachlor data collected for the LMMBP); LMI0040 (surficial sediment and sediment trap congener, total PCBs, and *trans*-nonachlor data); and LMI0041 (daily tributary congener, total PCBs, and *trans*-nonachlor loading data from 11 monitored and 18 unmonitored Lake Michigan tributaries).

At the same time, information about data received (metadata) was stored in a searchable Microsoft Access database. The database is found on the LLRS common drive "\\giord2\grlcommon", which is also known as the "r:\" drive. This database is named "lmtrack2000.mdb" and is found in the r:\access2000\ folder. This database is available to all staff. This database can be searched by library number (consecutive number assigned when data are logged in, corresponds to LMI folder name in lmmbdata archive), PI, parameter, PI and parameter, or library number and parameter (Appendix 1.3.3.).

After initial review of a data set was completed, data were retrieved from the Microsoft Access databases and exported into files (usually Microsoft Excel) for assessment by the modeler who would be using the data set. Water and sediment PCB data were given to Xiaomi Zhang. Lower food chain data were assessed by Xin Zhang and Katie Taunt. Forage and predator fish data were assessed by Xin Zhang. Initially, only routine field samples and field duplicates were given to the data assessors. If issues or problems were found, the person assessing the data would then request additional QA data. If questions/problems could not be resolved by looking at QA data, they were referred back to GLNPO for resolution.

In several instances, data which passed GLNPO QA checks from the analytical point of view were rejected during assessment because values were not environmentally reasonable. For example, particulate and dissolved water PCB values from station MB63 from the October 1995 cruise were orders of magnitude higher than values for surrounding stations. In addition, their values were orders of magnitude higher than values from the same station collected on different cruises. Environmentally these results were unreasonable, and they were not used by the modelers. GLNPO was informed whenever we rejected data.

After the assessment process was completed, files were created which could be used in IDL, which is a software package used for visualization and analysis of LMMBP data. Standard formats were developed for water, sediment, and fish data (Appendices 1.3.4, 1.3.5, 1.3.6). All files were fixed format ASCII text files. One of the principal uses of IDL was to develop volume-weighted averages (VWA) estimates of parameter concentrations for each cell in the modeling grid. These VWA estimates could then be compared to model results.

1.3.2 Calculation of Total PCBs

In general, total PCBs were calculated by the PI reporting the data. In the case of tributary loads for total PCBs, total PCBs were calculated by the PI, and loads were calculated by David Hall, USGS. In a similar fashion, total PCBs were calculated by the PI, and atmospheric loads were calculated by Keri Hornbuckle, University of Iowa. Open lake total PCBs were calculated by GLNPO contractor staff (Marcia Kuehl). A GLNPO contractor, DynCorp, verified total PCB values. However, the method used to calculate total PCBs was not consistent from PI-to-PI. Some PIs blanked corrected data; some included invalid data (samples with INV analytical remark field); and some did not surrogate correct data. In those instances when invalid samples were included in the total PCB calculation or surrogate correction was not done, total PCBs were recalculated by DynCorp to correct these problems. Attached are documents from Marcia Kuehl, PCBs QA Coordinator, describing how total PCBs were calculated by each PI (Appendix 1.3.7).

1.3.3 Regression Analysis of Measured Congener, Total PCB Data

As the modeling study was originally devised, all modeling was to be done at the congener level; however, at a later date it was decided that simulation of total PCBs would also be desirable. The Level 2 (LM2) and Level 3 (LM3) LMMBP models did not model total PCBs; therefore, a method was devised to calculate total PCB concentrations for model results based on the set of congeners modeled. Regressions and ratios were calculated comparing the PIs' measured total PCB field values (the independent variable) to the PIs' measured sum of the congeners that were modeled at the LLRS (the dependent variable) in all media modeled (atmospheric vapor phase, wet and dry deposition, dissolved and particulate tributary water, total, dissolved and particulate water, surficial sediment, phytoplankton, zooplankton, *Diporeia, Mysis,* and all forage and predator fish species). Note that total PCBs in water were not measured, but were derived by adding up dissolved and particulate PCBs for each sample. With R**2 values of .90 or greater, these regression analyses produced very good results (Table 1.3.1).

Additional analysis was then performed to produce an uncertainty estimate for the regression equations. A mean was calculated for the slope of the line in the linear regression (z), and 95 percent confidence intervals were calculated for z using the formula:

$z = x/y$

where

x = total of PCB congener subset that was modeled

y = true total PCBs as calculated by the PI.

David Miller, statistician at LLRS, verified that the z values were generally normally distributed. This allowed us to calculate a mean, standard deviation,

Table 1.3.1. Revised Regression Equations for the LMMBP Total PCBs in All Media

Media	Ratio of PI Calculated Total PCBs to Summed Modeling Congeners	PIs' Calculated Total PCBs Versus Summed Modeling Congeners: Regression Equation	R^2
Atmospheric Vapor Phase	1.2944	y = 1.2707x + 0.0891	0.9997
Atmospheric Dry Deposition	1.3597	y = 1.3204x + 0.2159	0.9623
Atmospheric Wet Deposition	1.5775	y = 1.6917x - 0.0322	0.9672
Tributary Loading Data	1.2476	y = 1.2134x + 0.7752	0.991
Dissolved Water	1.4822	y = 1.2738x + 0.0268	0.9413
Particulate Water	1.2948	y = 1.2251x + 0.0051	0.9992
Dissolved + Particulate Water	1.4147	y = 1.2427x + 0.0347	0.9829
Surficial Sediment	1.1805	y = 1.1668x + 0.6125	0.997
Phytoplankton	1.3842	y = 1.2871x + 3.6216	0.9584
Zooplankton	1.3923	y = 1.2058x + 22.833	0.9595
Diporeia	1.3652	y = 1.3763x - 3.4124	0.9795
Mysis	1.3162	y = 1.3829x - 12.842	0.9833
Alewife < 120 mm	1.4458	y = 1.4534x - 1.296	0.9784
Alewife > 120 mm	1.4281	y = 1.3338x + 38.145	0.947
Bloater < 160 mm	1.4761	y = 1.4317x + 19.505	0.9851
Bloater > 160 mm	1.4827	y = 1.4146x + 38.14	0.9426
Deepwater Sculpin	1.5157	y = 1.3752x + 38.735	0.9897
Slimy Sculpin	1.4976	y = 1.5272x - 8.4009	0.9257
Adult Smelt	1.4447	y = 1.46693x - 5.2828	0.9044
Hatchery Coho	1.2836	y = 1.4009x - 11.024	0.994
Coho Yearling	1.497	y = 1.6263x - 16.984	0.9835
Coho Adult	1.444	y = 1.4392x + 2.7179	0.9927
Adult Lake Trout	1.4897	y = 1.4875x + 3.7424	0.9977

and 95 percent confidence intervals for the z value in all media modeled (Appendix 1.3.8).

A comparison of field measured water congeners values to model results from Xiaomi Zhang, WelSo modeler at LLRS, indicated that there were some problems with measured field results for congeners 84+92 and 99. In both dissolved and particulate water fractions, concentrations for both of these congeners were much higher than model results.

Investigation into this issue revealed a contamination issue with congener 99. The analytical technique used to measure the water congener could not adequately separate congener 99 from *trans*-nonachlor, which caused a co-elution problem. The reason for the high field values for congener 84+92 was unclear, but it also is believed to be caused by a co-elution problem. Xiaomi Zhang developed ratios comparing field data to model results (Table 1.3.2).

These ratios were used to correct 84+92 and 99 congener values by the following formula:

Corrected congener nn value = Measured congener nn value/ratio

where

nn = congener 84+92 or 99.

Table 1.3.2. Ratio of Measured Field Data/Model Results for Congeners 84+92 and 99 in Water

Congener	Ratio in Dissolved Water	Ratio in Particulate Water
84+92	2.08	3.41
99	1.15	1.94

Regression analysis was then redone for dissolved, particulate, and dissolved + particulate water.

These revised regression equations were then applied to summed modeled congeners to calculate modeled total PCBs. Regression equations for dissolved and particulate fractions of total PCBs have a positive y-intercept. This means that when these regressions are used to calculate total PCBs, the value will never drop to zero even if modeled congeners drop to zero. Meetings were held with the modeling staff to discuss this issue. It was believed that this bias was caused by 1) lack of blank correction of congener data, 2) detection limits, and 3) inherent uncertainty in the regression process. Since it was not possible to correct these problems, the decision was made to use the regression equations, and carefully explain these difficulties when documenting modeling scenarios.

1.3.4 Summary

The LMMBP data received at the LLRS were carefully evaluated prior to use to insure that the field data being used by the modelers were as accurate as possible. In addition, data were archived and cataloged to protect these valuable data sets and make it easier for users to find the information. Incorporation of this information into LLRS Microsoft Access databases has given us flexibility in retrieving the information needed by the modeling staff at the LLRS. These data were used to develop regression equations which were used to approximate total PCB concentrations for modeled data.

References

Sukloff, W.B., S. Allan, and K. Ward. 1995. RDMQ User Manual. Environment Canada, Atmospheric Environment Service, North York, Ontario, Canada. 91 pp.